Journal of Mathematical Chemistry, Vol. 40, No. 4, November 2006 (© 2006)
DOI: 10.1007/s10910-006-9058-0

Density-based globally convergent trust-region methods for self-consistent field electronic structure calculations

Juliano B. Francisco

Department of Applied Mathematics, IMECC-UNICAMP, State University of Campinas, Campinas, SP, Brazil
E-mail: juliano@ime.unicamp.br

José Mario Martínez*

Department of Applied Mathematics, IMECC-UNICAMP, State University of Campinas, CP 6065, 13081-970 Campinas, SP, Brazil
E-mail: martinez@ime.unicamp.br

Leandro Martínez

Department of Physical Chemistry, IQ-UNICAMP, State University of Campinas, Campinas, SP, Brazil.
E-mail: lmartinez@iqm.unicamp.br

Received 15 October 2005; revised 8 December 2005

A theory of globally convergent trust-region methods for self-consistent field electronic structure calculations that use the density matrices as variables is developed. The optimization is performed by means of sequential global minimizations of a quadratic model of the true energy. The global minimization of this quadratic model, subject to the idempotency of the density matrix and the rank constraint, coincides with the fixed-point iteration. We prove that the global minimization of this quadratic model subject to the restrictions and smaller trust regions corresponds to the solution of level-shifted equations. The precise implementation of algorithms leading to global convergence is stated and a proof of global convergence is provided. Numerical experiments confirm theoretical predictions and practical convergence is obtained for difficult cases, even if their geometries are highly distorted. The reduction of the trust region is performed by a strategy that uses the structure of the energy function providing the algorithm with a nice practical behavior. This framework may be applied to any problem with idempotency constraints and for which the derivative of the objective function is a symmetric matrix. Therefore, application to calculations based both on Hartree–Fock or Kohn–Sham density functional theory are straightforward.

KEY WORDS: Hartree–Fock, Kohn–Sham, density functional theory, trust-region algorithms, Levenberg–Marquardt, convergence

*Corresponding author.

1. Introduction

Recent developments on the mathematical theory of self-consistent field (SCF) electronic structure calculations have provided algorithms of increased robustness [1–6]. Robustness, efficiency and user-independence are required for SCF calculations to become routinely used for obtaining the electronic structure of complex molecular systems, and also for the algorithms to be used safely by non-experts.

The first and basic algorithm considered for solving SCF equations was the well known fixed-point method [7]. This method converges only for very well behaved cases and it was soon realized that it could not be trusted for solving the Hartree–Fock (HF) equations of most molecular systems. Several approaches were developed in order to obtain more reliable algorithms: The classical Newton and conjugate gradient optimization methods were applied to parametrized versions of the SCF equations [8,9] and very successful accelerations, particularly the DIIS method of Pulay [10,11], provided greater stability for the fixed-point iteration. More recently, Cancès and Le Bris used the structure of the HF equations in an very elegant manner and proposed the optimal damping algorithm (ODA), for which they proved global convergence whenever the iterates satisfy the Uniform-Well Posedness assumption [1,2]. The ODA was observed to have a slow convergence and a combination of ODA with DIIS called EDIIS was proposed by Kudin et al. [12] for which increased robustness is obtained relative to DIIS. The successful implementation of EDIIS led it to be implemented in the Gaussian package [13].

Also recently, Thogersen et al. and the authors of the present work started working on the application of trust-region strategies for the SCF problem [3,5,14]. In our previous work, we defined a general globally convergent algorithm applicable to the SCF equations. A theoretical framework for this algorithm and its variations was developed and global convergence was proved without any assumption on the sequence of iterates [5]. Our approach was based on the minimization of a quadratic approximation of the energy as is usually expected from a trust-region algorithm. We noted that, although convergence was achieved in all problems, as expected, it was slow. That was basically because we used a very crude quadratic approximation that was based on the energy as a function of the coefficient matrix. Thogersen et al. [3,4] on the other hand, addressed the problem of minimizing the energy as function of the density matrix originally in HF and more recently Kohn–Sham density functional theory SCF calculations. Since the energy in HF equations is quadratic as a function of the density matrix, it is expected the optimization algorithms to behave better when the variable is the Density. Indeed, they have shown a very nice practical behavior when their trust-region based algorithm was coupled with also trust-region-based accelerations. Their approach is very interesting, but it is not a rigorous application of a full-featured trust-region method.

In this paper we develop the theory and the basic algorithm required for a full-featured, globally convergent algorithm for SCF calculations based on density matrices. Our approach has a close correspondence with the work of Thogersen et al., but the precise use of the trust-region structure provides the algorithm with global convergence properties. In the development of our particular implementation of the method, we also use the structure of the HF equations, employed originally by the ODA algorithm, resulting on a density-based trust-region method with very nice practical behavior even without accelerations.

This paper is organized as follows. In section 2 we overview the fundamental properties of globally convergent trust-region methods. In section 3 we define the optimization problems that we aim to solve using the trust-region approach. This class of problems includes Hartree–Fock and Kohn–Sham SCF calculations. In section 4 we describe the quadratic subproblems and we prove a theorem that gives their exact *global* solution. In section 5 we define the density-based globally convergent trust-region method (DGTR). In section 6 we discuss the arguments that lead to global convergence. In section 7 we describe the choice of the trust-region size (or, equivalently, Levenberg–Marquardt parameter). In section 8 we show numerical experiments and discuss how accelerations or non-monotone strategies may be coupled with DGTR methods. Conclusions are stated in section 9. The appendix A contains the rigorous proof of the main theorem in section 4.

2. Fundamentals of trust-region methods

The classical constrained optimization problem consists in the minimization of a scalar multivariate function subject to constraints which, in general, are described by equalities and inequalities. Trust-region methods for unconstrained optimization were introduced by Powell in 1970 [15,16]. Powerful convergence properties were proved by Sorensen in 1982 [17]. Trust-region ideas were used in sequential quadratic algorithms (SQP) for constrained optimization [15]. At each iteration of SQP, the feasible set is approximated by a polyhedron and the objective function is modeled by a quadratic. Only in 1995, Martínez and Santos [18] proposed a trust-region method for constrained optimization where the feasible set is not approximated at all by its linearization. As most numerical algorithms for non-linear optimization, the trust-region method [18] is iterative and generates a sequence of feasible points such that the objective function decreases monotonically. Given the current iterate, the steps that lead to a better approximation to the solution may be sketched as follows.

1. Define a quadratic model for the objective function.
2. Minimize the quadratic model on the intersection of the feasible set and the trust region. The minimizer of the quadratic so far obtained is called

"trial point". The difference between the quadratic function values at the current point and at the trial point is called "Predicted reduction" (**Pred**).

3. The "Actual reduction" (**Ared**) is the difference between the objective function values at the current point and at the trial point. If **Ared** is large enough when compared with **Pred**, the trial point is "accepted". If this is not the case, the radius of the trust region is reduced and the quadratic minimization step is repeated.

4. In the case of acceptance of the trial point, the new iterate is chosen as any feasible point such that the objective function value is less than or equal to the one at the accepted trial point.

The quadratic model must be an approximation of the true objective function. This requires that the gradient of the quadratic model must be the gradient of the objective function at the current point. It is desirable that the second derivatives of the quadratic model also coincide with the second derivatives of the objective function at the current point but, many times, this is not possible. In any case, the gradient agreement guarantees that, close to a non-critical current point, the decrease of the quadratic model implies the sufficient decrease of the objective function. Therefore, after a finite number of reductions of the trust-region radius, the trial point is necessarily accepted.

A critical feasible point is a point that satisfies some optimality condition of non-linear optimization. Global convergence results usually say that limit points generated by optimization algorithms are critical. In the case of the trust-region framework, for getting global convergence it is essential that the trial point satisfies a sufficient descent condition (**Ared** must be at least a fraction of **Pred**) instead of a single descent condition (**Ared** positive). Moreover, in the trust-region convergence theory the trust regions are "balls" (set of points whose "distance" to the current point is less than or equal to a given trust-region radius) although different definitions of distance may be employed.

The accepted trial point is not necessarily the point adopted as new iterate. Very frequently one tries something "even better", and the only requirement for this "accelerated" point is the decrease of the objective function value with respect to the trial point. Most times, this acceleration device is not mentioned in global convergence theories which, on the other hand, hold straightforwardly employing the acceleration. In practical implementations, however, acceleration steps are very popular [19].

The main drawback for the implementation of the trust-region method on arbitrary domains [18] sketched above is that the subproblem of minimizing the quadratic on the intersection of the feasible set and the trust region may be very difficult. A "Levenberg–Marquardt"-like modification of the basic algorithm [5,14] removes, partially, this drawback. Instead of dealing directly with

the trust region, a penalty term is added to the quadratic objective function of the subproblem and the trust-region constraint is eliminated. The penalty term is proportional to the (squared) distance to the current point and depends on a penalty parameter that is increased when the actual reduction fails to be sufficiently larger than the predicted reduction. It may be proved [5,14] that the effect of increasing the penalty parameter is the same as the effect of decreasing the trust-region radius. Roughly speaking, this implies that the convergence properties of the Levenberg–Marquardt-like modification are the same as the convergence properties of the pure trust-region method [18].

Thanks to the Levenberg–Marquardt modification, the trust-region method becomes implementable in many cases in which the practical solution of the subproblem seems to be impossible. However, the applicability of the method is still subject to the solvability of a quadratic subproblem on the feasible region.

3. Optimization problems in SCF electronic structure calculations

Here we give precise definitions of the optimization problems studied. We first provide a very general definition in order to stimulate theoretical work from the optimization community. Next, we define the particular case of closed-shell restricted HF equations that are explicitly studied in this paper. We note, however, that all the methods presented here could well be applied to any problem that fit into the general formulation, as do electronic structure calculations based on Kohn–Sham density functional theory.

3.1. General formulation

The general problem that we have in mind is

$$\text{Minimize } E(D) \text{ subject to } D \in \mathcal{M}, \tag{1}$$

where

$$\mathcal{M} = \{D \in I\!R^{K\times K} \mid DSD = D,\, D^T = D,\, \mathrm{Tr}(DS) = N\} \tag{2}$$

and S is symmetric and positive definite. We assume that $\nabla E(D)$ is a symmetric $K \times K$ matrix.

It is generally accepted that the global minimizer of electronic structure calculations must be *aufbau*. This has a precise meaning for HF equations related to the construction of the density matrix from the set of eigenvectors of the Fock matrix corresponding to N lowest eigenvalues (see below). A proof for infinite-dimension, unrestricted HF equations that the global minimizer is *aufbau* was given by Lions [20]. His proof cannot be straightforwardly adapted to other problems.

The theoretical framework of the algorithm presented here led us to define an *aufbau* point as a *global minimizer of the linear approximation of the objective function in the feasible set.* We believe that this general definition, which is applicable to any optimization problem, could be valuable for further efforts in order to solve the following questions in more general terms: Which should the properties of the objective function and the constraints be for the global minimizers be *aufbau*? Is it possible to develop an algorithm that converges necessarily to *aufbau* solutions? The answer of both these questions would be of great value for the theoretical comprehension of the electronic structure optimization problem.

3.2. Closed-shell restricted HF equations

The optimization problem in closed-shell restricted HF equations can be expressed in the following way [7]:

$$\text{Minimize } E_{\text{SCF}}(D) \text{ subject to } D \in \mathcal{M}, \tag{3}$$

where

$$E_{\text{SCF}}(D) = \text{Tr}\Big[2HD + G(D)D\Big].$$

D is the one-electron density matrix in the atomic-orbital basis, H is the one-electron Hamiltonian matrix, $G(D)$ is given by

$$G_{\mu,\nu}(D) = \sum_{\rho=1}^{K}\sum_{\sigma=1}^{K}(2g_{\mu\nu\rho\sigma} - g_{\mu\sigma\rho\nu})D_{\sigma\rho},$$

$g_{\mu\nu\rho\sigma}$ is a two-electron integral in the AO basis, K is the number of functions in the basis and $2N$ is the number of electrons.

Clearly [7]:

$$\nabla E_{\text{SCF}}(D) = 2F(D),$$

where

$$F(D) \equiv H + G(D).$$

$F(D)$ is known as the *Fock Matrix*. Since $G(D)$ is linear, the function $E_{\text{SCF}}(D)$ is quadratic.

The matrices D that belong to $\mathcal{M}$ may be written as $D = CC^T$, where C is a real $K \times N$ matrix with S-orthonormal columns. That is, $C^T SC$ is the Identity $N \times N$ matrix. With this replacement, the problem (3) may be formulated in terms of the "Coefficients matrix" C. A Levenberg–Marquardt-trust-region method for solving this reformulation was introduced in [5]. Here we will concentrate ourselves on the Density-matrix formulation (3).

4. Quadratic subproblem

We aim to solve (1) using the trust-region framework and assuming that the current iterate is the matrix $\bar{D}$. We need a quadratic approximation of $E(D)$ with the condition that the gradient of the quadratic should be $\nabla E(\bar{D})$. The simplest form of a quadratic approximation is the linear approximation:

$$L_k(D) = E(\bar{D}) + \mathrm{Tr}\Big[\nabla E(\bar{D})(D - \bar{D})\Big]. \tag{4}$$

This approximation satisfies the requirement $\nabla L_k(\bar{D}) = \nabla E(\bar{D})$ and, thus, defines an admissible model.

According to the trust-region philosophy, the subproblem at iteration k should be:

$$\text{Minimize } L_k(D) \text{ subject to } D \in \mathcal{M}, \|D - \bar{D}\| \leqslant \Delta, \tag{5}$$

where Δ is the trust-region radius and $\|.\|$ represents a norm in the space of $K \times K$ matrices, which will be specified later.

Thogersen et al. [3] studied the application of this framework to the SCF problem and to the Kohn–Sham problem [4]. They chose:

$$\|A\| = \|A\|_S = \sqrt{\mathrm{Tr}(ASAS)} = \|S^{1/2}AS^{1/2}\|_F,$$

where $\|\cdot\|_F$ is the Frobenius norm ($\|A\|_F^2 = \sum_i \sum_j A_{ij}^2$). They observed (equation (16) of [4]) that, when $D \in \mathcal{M}$,

$$\|D - \bar{D}\|^2 = -2\mathrm{Tr}(\bar{D}SDS) + 2N.$$

Accordingly, they replaced the trust-region constraint $\|D - \bar{D}\| \leqslant \Delta$ by the linear equation

$$-2\mathrm{Tr}(\bar{D}SDS) + 2N = \Delta^2$$

and deduced the level-shifted [21] Roothan–Hall equations using the Lagrange conditions of the modified form of (5). Our Theorem 1 below may be interpreted as a rigorous justification of this procedure. The full rigorous justification is necessary because the Lagrange conditions provide only necessary conditions for local minimization and in the trust-region approach we wish the *global* solution of (5).

The choice of $\|\cdot\| = \|\cdot\|_S$ in (5) is crucial because, otherwise, subproblem (5) should be very difficult (or impossible) to solve accurately. Using the Levenberg–Marquardt approach, we may eliminate the trust-region constraint $\|D - \bar{D}\| \leqslant \Delta$ adding a penalty term to the objective function. In this way, we arrive to the subproblem:

$$\text{Minimize } L_k(D) + \mu\|D - \bar{D}\|^2 \text{ subject to } D \in \mathcal{M}. \tag{6}$$

Here, $\mu > 0$ is the penalty parameter associated with the trust-region constraint. The effect of $\Delta \to 0$ is essentially the same of $\mu \to \infty$ [5,14].

By (4), the subproblem (6) is equivalent to:

$$\text{Minimize } \mathrm{Tr}\left[\nabla E(\bar{D})(D-\bar{D})\right]+\mu\mathrm{Tr}\left[(D-\bar{D})S(D-\bar{D})S\right] \text{subject to } D\in\mathcal{M}. \quad (7)$$

The following theorem gives the exact *global* solution of (7). Its proof is given in the appendix A.

Theorem 1. Let S, W and $\bar{D}$ be symmetric $K \times K$ real matrices. Assume that S is positive definite. Let $V = (V_1, \ldots, V_N) \in \mathbb{R}^{K\times N}$ be a matrix whose columns $V_1, \ldots, V_N$ are generalized eigenvectors corresponding to the N smaller generalized eigenvalues of

$$W - \mu S\bar{D}S.$$

(This means that $\lambda_1, \ldots, \lambda_N$ are the smaller numbers that satisfy

$$\left[W - \mu S\bar{D}S\right]V_i = \lambda_i S V_i, \quad i = 1, \ldots, N, \quad (8)$$

and

$$V_i^T S V_j = \delta_{i,j} \; \forall \; i \neq j. \quad (9)$$

Let $D_{\text{trial}} = VV^T$. Then, D_{trial} is a solution of

$$\text{Minimize } \mathrm{Tr}\left[2W(D-\bar{D})\right]+\mu\mathrm{Tr}\left[(D-\bar{D})S(D-\bar{D})S\right], \quad \text{subject to } D\in\mathcal{M}. \quad (10)$$

Moreover, the optimum value of (10) is $2(\lambda_1 + \cdots + \lambda_N) + c$, where

$$c = \mu[N + \mathrm{Tr}[\bar{D}S\bar{D}S] - 2\mathrm{Tr}[W\bar{D}].$$

Taking $W = \nabla E(\bar{D})/2$, theorem 1 shows that the trust-region subproblem (7) is solvable and so, it can be the basis of a globally convergent Levenberg–Marquardt trust-region algorithm. Observe that the assumption $\bar{D} \in \mathcal{M}$ is not necessary in theorem 1.

Theorem 1 represents a rigorous justification for the fact that level-shifted Roothan–Hall equations provide global solutions of the trust-region subproblem (5).

Since theorem 1 is independent of the choice of W, it turns out that it may be applied to the resolution of trust-region subproblems arising from any optimization problem where the feasible set is $\mathcal{M}$, given that the gradient of the objective function is a symmetric matrix.

If $\mu = 0$, the solution of (7) is a global minimizer of the linear approximation of E_{SCF} on the feasible region $\mathcal{M}$. This solution corresponds to the classical fixed-point iteration for solving the SCF problem.

5. Globally convergent trust-region method

Although we have in mind the SCF problem, the algorithm below apply to any optimization problem of the form (1) to (2), given that, for all D, the gradient $\nabla E(D)$ is a symmetric $K \times K$ matrix. Moreover, the symmetry of $\nabla E(D)$ is necessary only to show that the global solution of the trust-region subproblem is as given by theorem 1.

Algorithm 1

Let $\alpha \in (0, 1/2)$, $\mu_{\max} > 0$, $1 < \tau_{\min} < \tau_{\max} < \infty$.

Step 1. Choose $D_0 \in \mathcal{M}$ and set $k \leftarrow 0$.

Step 2. Choose $\mu_{\mathrm{first}} \in [0, \mu_{\max}]$ and set $\mu \leftarrow \mu_{\mathrm{first}}$.

Step 3. Solve (7), using (8), with $\bar{D} = D_k$ and obtaining the solution $D_{\mathrm{trial}} (= D_{\mathrm{trial}}(\mu))$.

Define

$$\mathbf{Pred} = L_k(\bar{D}) - L_k(D_{\mathrm{trial}}) = E(\bar{D}) - L_k(D_{\mathrm{trial}}) = \mathrm{Tr}[E(\bar{D})(\bar{D} - D)].$$

If $\mathbf{Pred} = 0$, terminate the execution of the algorithm.

Step 5. Define

$$\mathbf{Ared} = E(\bar{D}) - E(D_{\mathrm{trial}}).$$

If

$$\mathbf{Ared} \geqslant \alpha \mathbf{Pred}, \tag{11}$$

compute $D_{k+1} \in \mathcal{M}$ (acceleration) such that

$$E(D_{k+1}) \leqslant E(D_{\mathrm{trial}}), \tag{12}$$

set $k \leftarrow k + 1$ and go to Step 2.

If (11) does not hold, then, if $\mu = 0$ take $\mu_{\mathrm{new}} > 0$. If $\mu > 0$, take $\mu_{\mathrm{new}} \in [\tau_{\min}\mu, \tau_{\max}\mu]$. Set $\mu \leftarrow \mu_{\mathrm{new}}$ and go to Step 3. □

6. Convergence proof arguments

The global convergence theory of [5,14,18] guarantees that limit points generated by Algorithm 1 are stationary (or critical). More specifically, this algorithm is a particular case of Algorithm B2 of [5]. A small technical difficulty must be removed. In the theory of [5] it is assumed that all the feasible points are regular, a property that is not true if $\mathcal{M}$ is described in the form (2). However, $\mathcal{M}$ can also be described as the set of $K \times K$ matrices such that $D = XX^T$, with X being a $K \times N$ matrices with S-orthonormal columns. In this way it is easy to see that the set of pairs (D, X) that satisfy that property contains only regular points. Therefore, the regularity argument invoked in the proof of Lemma B.2 of [5] remains valid. (Regularity is a property that depends of the representation of the constraint set, but not of the constraint set itself.)

Let us review here the main features that imply global convergence of the trust-region algorithm.

1. *The sufficient descent condition.* In (11) we require that the energy at the new point is, not only smaller, but *sufficiently smaller* than the energy at the current point D_k. Single descent is not enough for proving global convergence. The reason is that, with a single descent condition, we could obtain a sequence $E(D_0) > E(D_1) > E(D_2) \ldots$ that approximates monotonically to a non-stationary point $\hat{D}$. For example, if $E(D_k) > E(D_{k+1}) \geqslant E(D_k) - \varepsilon/2^k$ for all k, we would have $E(\hat{D}) \geqslant E(D_0) - \varepsilon$. The value of α measures the desired degree of sufficient descent. There are rival claims about the sensible value for α both in trust-region as in line-search optimization. A small value of α allows one to take larger steps but, on the other hand, the descent could be unsatisfactory. In trust-region methods $\alpha = 0.1$ is generally used, whereas in line-search algorithm, $\alpha = 10^{-4}$ is usually preferred.

2. *Safeguarding parameter* $\mu_{\max}$. The *initial* penalty parameter employed at iteration k of the algorithm must be smaller than a given number $\mu_{\max}$. This means that the initial trust-region radius used at iteration k should not be very small. The reason is obvious: if one admits smaller and smaller initial trust-region radii at different iterations, the sequence of iterates D_k could converge, with monotonic decrease of the energy, to a non-stationary point $\hat{D}$. For example, if $0 < \|D_{k+1} - D_k\| \leqslant \varepsilon/2^k$ we would have $\|D_0 - \hat{D}\| \leqslant \varepsilon$, a restrictive property that, in general, is not true at stationary points.

   The argument above applies to the *initial* penalty parameter at iteration k but not to the *finally accepted* one. In fact, we may finish the iteration accepting $\mu > \mu_{\max}$, but only after failures of the sufficient descent condition (11). In other words, it is admissible a very small trust region, if a larger one does not provide the necessary decrease, but one

should always begin the iteration trying a "not very small" trust region radius.

3. *Penalty-increasing parameters*. When D_{trial} does not satisfy the sufficient descent condition (11) the penalty parameter μ must be increased (the trust region must be decreased). The increase of μ must be controlled by two parameters $\tau_{\min}$ and $\tau_{\max}$. The minimum increase is controlled by $\tau_{\min} > 1$. This guarantees that, eventually, the penalty parameter at iteration k is large enough, so that the sufficient decrease condition is met. On the other hand, we cannot increase the penalty parameter very abruptly and, for this reason, we impose $\mu_{\text{new}} \leqslant \tau_{\max}\mu_{\text{old}}$. An abrupt increase of μ corresponds to an abrupt decrease of Δ and its effect could be the same as the effect of beginning the iteration with excessively large values of μ.

It is worthwhile to stress that the fact that all the limit points are stationary [5] only depend of the features mentioned above. In (12) the method is allowed to choose a new point that could be even better than the finally accepted point in the main block of the algorithm. Convergence proofs do not depend at all on this choice, which, on the other hand, could be very important, or even essential, for the efficiency of the algorithm.

7. Choice of the trust-region size

7.1. TRRH method

The TRRH and TRSCF methods of Thogersen et al. [3,4] may be analyzed under the framework of algorithm 1. The similarities are:

- The trial points are obtained using the subproblem (8). As we mentioned in the previous section, Thogersen et al. [3] arrive to this problem considering the Lagrange equations associated with the boundary trust-region subproblem. In theorem 1 we proved that, in fact, this solution gives the *global* minimizer of the trust-region subproblem.
- A clever acceleration procedure is employed in [3,4] for obtaining a better density matrix, as in (12). Their acceleration procedure is based on the resolution of trust-region subproblems in an auxiliary space, generated by the density matrices at previous iterations.

The main difference between algorithm 1 and the approach of [3,4] is that Thogersen et al. [3,4] do not use the sufficient decrease condition (11) at all. Instead, they require:

$$\|D_{\text{trial}}(\mu) - \bar{D}\|_S \leqslant 0.2\sqrt{N}, \tag{13}$$

which is equivalent to the condition $a_{\min} \geqslant 0.98$ in [3]. Since $\|\bar{D}\|_S = \sqrt{N}$, condition (13) means that the difference $\|\bar{D} - D_{\text{trial}}\|_S$ is less than 20% the norm of $\bar{D}$.

In their implementation, Thogersen et al. [4] consider

$$D_{\text{trial}}(\mu) = C_{\text{trial}}(\mu) C_{\text{trial}}(\mu)^T, \bar{D} = \bar{C}\bar{C}^T,$$

where C_{trial} and $\bar{C}$ are $S-$orthonormal $K \times N$ matrices,

$$\bar{C} = (\bar{\phi}_1, \ldots, \bar{\phi}_N), C_{\text{trial}} = (\phi_1^{\text{trial}}, \ldots, \phi_N^{\text{trial}}).$$

They define, for all $i = 1, \ldots, N$,

$$a_i^{\text{orb}} = \sum_{j=1}^{N} (\bar{\phi}_j^T S \phi_i^{\text{trial}})^2,$$

and they require:

$$\min_i a_i^{orb} \geqslant 0.98. \tag{14}$$

It is easy to see that (14) implies (13).

Their procedure, may be schematized as follows [Lea Thogersen, private communication [22]]:

1. If $D_{\text{trial}}(0)$ satisfies (14), accept this D_{trial} and go to the acceleration phase.
2. Otherwise, take μ_{left} as the last number in the sequence $\{0, 10, 15, 20, 25, \ldots\}$ that *does not* satisfy (14) and μ_{right} as the first number of this sequence that satisfies (13).
3. Use a bisection procedure in the interval $[\mu_{\text{left}}, \mu_{\text{right}}]$ for obtaining $\mu \in [\mu_{\text{left}}, \mu_{\text{right}}]$ satisfying (14) and such that $\mu - 0.1$ does not satisfy (14). In other words, the equality should hold in (14) with precision 0.1.

Thogersen et al. [3,4] observe that, with this choice of μ, the energy frequently decreases and the HOMO-LUMO gap remains positive. However, energy decrease is not always verified, as the numerical experiments will show. In order to guarantee the decrease in energy at every iteration, the parameter μ must be increased (the trust region must be reduced) every time that an increase in energy is obtained, and a new trial point must be computed. This is not done by the TRRH strategy, which considers a trust region of constant size.

As the numerical experiments presented below show, an algorithm with a fixed-size trust region is well behaved only far from the solution. We illustrate the reason for this behavior by a simple example: The minimization of single variable quadratic function. Recall that these methods are based on the global

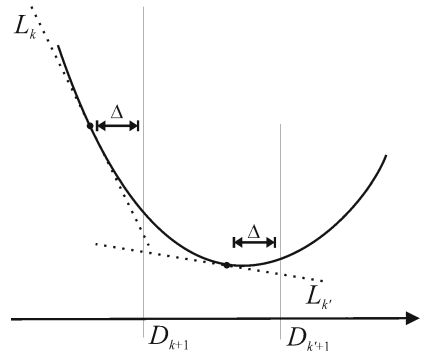

Figure 1. The effect of a trust region of constant size Δ in iterations far from a minimum (k) and close to the solution (k'). The minimization of the linear approximation in this trust region provides a sufficient decrease of the quadratic function at iteration k, but results in an increase in the value of the quadratic in iteration k'. The point $D_{k'+1}$ should not be accepted in order to achieve convergence.

minimization of the linear approximation of the true energy at every iteration. As can be seen in figure 1, when far from the solution the global minimization of the linear approximation provides points with lower energy. However, when the derivative of the objective function is small, the minimization of the linear approximation may provide a new point that is on the other side of the minimum and that may have a higher energy. Even worse, if this point higher in energy is accepted, the minimization of the linear approximation with the same trust-region size might provide again the same point as before and the algorithm will oscillate between these two points rather than converge to the minimum. This problem does not occur when points higher in energy are rejected and the trust region is reduced, forcing the convergence to the energy minimum. In the work of Thogersen et al. [3,4] this problem is overcome by the introduction of very clever acceleration strategies that are also based on trust-region arguments. The accelerations also provide the algorithms with high efficiency, and we will briefly describe how they can be introduced in our DGTR methods without affecting global convergence properties. Their actual implementation is not the purpose of the current work.

7.2. Optimally damped DGTR

When Cancès and Le Bris first proposed the ODA, they used the interesting property of the HF energy that, given a line in the density matrix space, it is straightforward to compute the global minimizer of the unrestricted energy along that line [1,2,12]. We extend this idea a little further: Suppose that we have performed a fixed-point iteration that resulted in an increase in the energy. The fixed-point iteration is the minimization of the linear approximation of the

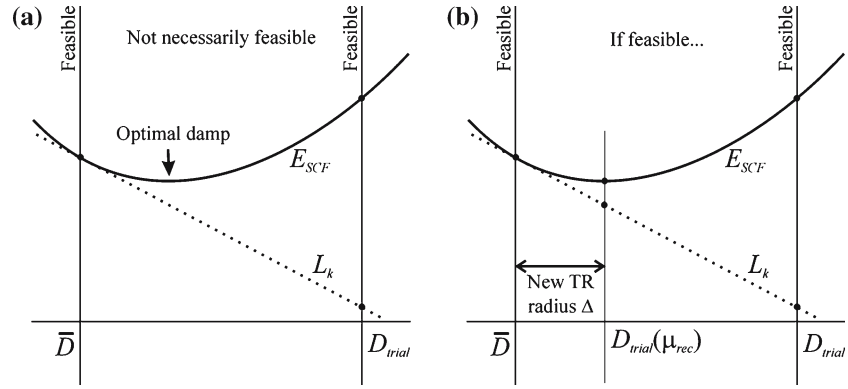

Figure 2. The optimally damped choice of μ: (a) The parabola joining the current point $\bar{D}$ and a unsuccessful trial point D_{trial} is convex and its minimum is in the segment $[\bar{D}, D_{\text{trial}}]$. The optimal damp would be its minimizer, but it is not necessarily feasible. (b) The optimally damped choice of μ is such that, if the segment were feasible, the new trial point would coincide with the minimizer of the energy along that segment.

energy, and the unrestricted energy function is quadratic as a function of the density matrix. Figure 2(a) shows that the unrestricted energy function must be convex in the line connecting the current point $\bar{D}$ and the rejected trial point D_{trial}. This is a consequence of the derivative of the energy being the same as the derivative of the linear approximation in $\bar{D}$ and of the energy (E_{SCF}) being higher in D_{trial}. The minimizer of the unrestricted E_{SCF} clearly must be in the interval $[\bar{D}, D_{\text{trial}}]$. The minimizer of E_{SCF} in the segment $[\bar{D}, D_{\text{trial}}]$ can be readily computed by the strategy of Cancès and Le Bris. Unfortunately, this minimizer is not necessarily feasible and, therefore, it cannot be taken as the next iterate of the DGTR algorithm (feasibility is necessary for all iterates in order that the algorithm fits in the framework of trust region strategies on arbitrary domains [5,18]).

Let us suppose, for a moment, that this minimizer is indeed feasible. An excellent choice for the new trust region would be the trust region for which the minimizer of the linear approximation coincides with the minimizer of E_{SCF}, as shown in figure 2(b). This trust region corresponds to a particular value of the μ parameter. This value of μ is the one chosen in our optimally damped DGTR algorithm. Accordingly to the DGTR scheme, the linear approximation of the restricted energy function is globally minimized in the trust region defined by this choice of μ. The decrease of the energy will be similar to the decrease expected by the minimization of the unrestricted energy in the segment $[\bar{D}, D_{\text{trial}}]$ whenever there is a feasible point in the vicinity of the minimizer of E_{SCF}.

The strategy can be precisely stated in the following way. Let us define the parametric function $\varphi(t)$ $(t \in [0, 1])$ by

$$\varphi(t) = E(\bar{D} + t(D_{\text{trial}} - \bar{D})).$$

Clearly, this function fits the conditions of having the same value and derivative of E in $\bar{D}$ and the same value of E in D_{trial},

$$\varphi(0) = E(\bar{D}), \varphi'(0) = \text{Tr}[\nabla E(\bar{D})(D_{\text{trial}} - \bar{D})], \varphi(1) = E(D_{\text{trial}}),$$

even if E is not quadratic. For all $t \in [0, 1]$ let $p(t)$ be the parabola defined by

$$p(0) = \varphi(0), p'(0) = \varphi'(0), p(1) = \varphi(1).$$

Since **Ared** $> \alpha$**Pred** one has that $p(1) \geqslant \alpha p'(0)$, therefore the parabola is convex as in figure 2.

From

$$p(t) = p(0) + p'(0)t + p''(0)t^2/2,$$

and the fact that $p''(t)$ is constant, we obtain that

$$\begin{aligned} p''(t) &= 2[p(1) - p(0) - p'(0)] \\ &= 2[\varphi(1) - \varphi(0) - \varphi'(0)] = 2(E(D_{\text{trial}}) - E(\bar{D}) - \text{Tr}[\nabla E(\bar{D})(D_{\text{trial}} - \bar{D})]) \end{aligned}$$

for all t. Since p is convex, we have that $p''(t) > 0$. The minimizer of this parabola is, therefore,

$$t_{\min} = -\frac{p'(0)}{p''(0)} = \frac{-\text{Tr}[\nabla E(\bar{D})(D_{\text{trial}} - \bar{D})]}{2(E(D_{\text{trial}}) - E(\bar{D}) - \text{Tr}[\nabla E(\bar{D})(D_{\text{trial}} - \bar{D})])},$$

and the corresponding density matrix D is:

$$D_{\min} = \bar{D} + t_{\min}(D_{\text{trial}} - \bar{D}).$$

Now, we consider the restriction of (6) to the line $[\bar{D}, D_{\text{trial}}]$:

$$\text{Minimize } L_k(D) + \mu \|D - \bar{D}\|_S^2 \text{ subject to } D = \bar{D} + t(D_{\text{trial}} - \bar{D}), \qquad (15)$$

and we seek the value of μ for which the solution of (15) is $D_{\min}$. It turns out that this μ is

$$\mu_{\text{rec}} = \frac{\varphi''(t)}{2\|D_{\text{trial}} - \bar{D}\|_S^2}.$$

Therefore, following the safeguards relative to the choice of μ_{new} in algorithm 1, we define:

$$\mu_{\text{new}} = \max\{\tau_{\min}\mu_{\text{old}}, \min\{\tau_{\max}\mu_{\text{old}}, \mu_{\text{rec}}\}\}.$$

The choice of μ_{rec} admits a different interpretation: the objective function of (6) with $\mu = \mu_{\text{rec}}$ turns out to be a quadratic approximation of $E(D)$ in the sense that $L_k(D) + \mu_{\text{rec}}\|D - \bar{D}\|_S^2$ coincides with the interpolating parabola along the line that joins $\bar{D}$ and D_{trial}.

The case of HF equations is especially interesting because, in this case, $E_{\mathrm{SCF}}(D)$ is a quadratic and, so, the interpolating parabola coincides exactly with $E_{\mathrm{SCF}}(D)$ along the line $[\bar{D}, D_{\mathrm{trial}}]$. In other words, $p(t) = \varphi(t)$ for all t. Therefore, in that case, we also have that $p'(1) = \varphi'(1)$. So, since

$$p'(1) = p'(0) + p''(0),$$

we get the following formula for μ_{rec}:

$$\mu_{\mathrm{rec}} = \frac{\mathrm{Tr}[(\nabla E(D_{\mathrm{trial}}) - \nabla E(\bar{D}))(D_{\mathrm{trial}} - \bar{D})]}{2\|D_{\mathrm{trial}} - \bar{D}\|_S^2}.$$

We note this strategy can also be used for Kohn–Sham density functional theory based equations, but in that case the parabola $p(t)$ will not coincide exactly with energy function because it is not quadratic [4,7]. In that case one would expect this strategy to be less effective, although a more detailed study would be required.

With regards to the choice of the first penalty parameter μ_{first}, the large-step alternative $\mu_{\mathrm{first}} = 0$ is interesting because, in this way, the trial point corresponds to the classical fixed-point iteration in SCF calculations.

This choice of μ has a close relationship with the well known spectral choice in optimization, where

$$\mu_{\mathrm{first}} = \max\{0, \min\{\mu_{\max}, \mu_{\mathrm{spec}}\}\},$$

and

$$\mu_{\mathrm{spec}} = \frac{\mathrm{Tr}[(\nabla E_{\mathrm{SCF}}(D_{k-1}) - \nabla E_{\mathrm{SCF}}(\bar{D}))(D_{k-1} - \bar{D})]}{2\|D_{k-1} - \bar{D}\|_S^2}. \tag{16}$$

When $\mu = \mu_{\mathrm{spec}}$ the quadratic in (6) *coincides* with the parabolic interpolation of the true objective function $E(D)$ along *all the line* determined by D_{k-1} and D_k. So, this quadratic is a nice approximation of the true objective function and its minimization as a part of the resolution process is well justified. For discussions and applications of the spectral parameter see [5, 23–29].

8. Numerical experiments

We have performed 12 numerical experiments in order to test the properties of the algorithm proposed here. These examples are divided in two sets. The first set is composed of three diatomic molecules (Cr_2, CrC and Rh_2), a tetrahedral anion (RhF_4^-), a Rhenium complex, and an ordered arrangement of litium and fluorine atoms (challenging case provided by K. N. Kudin, personal communication). All these examples were already used in previous publications to test the convergence of new algorithms or were observed to provide unusually difficult

Table 1
Geometries and basis used for the numerical examples.

Molecule	Geometry	Basis
Diatomic molecules	Bond length = 2.0 Å	STO-3G
Distorted diatomic molecules	Bond length = 10.0 Å	STO-3G
RhF_4^-	Tetrahedric, bond length = 2.57 Å	Ahlrichs VDZ and STO-3G on Rh
Distorted RhF_4^-	Tetrahedric, bond length = 5.0 Å	Ahlrichs VDZ and STO-3G on Rh
Rhenium complex	Given in Ref. [3]	Ahlrichs VDZ and STO-3G on Rh
Distorted rhenium complex	Coordinates on the example above are multiplied by 2.	Ahlrichs VDZ and STO-3G on Rh
Li_9F_9	See table A.1	STO-3G
Distorted Li_9F_9	Coordinates of the example above are multiplied by 2.	STO-3G

convergence for the algorithms here tested [3,9]. Well behaved examples are of no interested here: The convergence behavior of DGTR algorithms is exactly the same as pure fixed-point algorithm when the fixed-point iteration is able to provide a sufficient decrease of the energy function at every iteration. The second set of tests is formed by the same set of molecules, but with highly distorted geometries. Distorted geometries are recognized to cause convergence instabilities due to the introduction of degenerescences, and, therefore, they are a good test to the robustness of the DGTR algorithm. The basis sets used and the geometries used are described in table 1.

In our implementation of DIIS, extrapolation is used from the second iteration on. Therefore, the first extrapolation uses two residuals. The number of residual vectors used is increased in the subsequent iterations up to a maximum of 10. The parameters used in the OD-DGTR algorithm are $\alpha = 10^{-4}$, $\mu_{\text{first}} = 0$, $\tau_{\min} = 1.1$, $\tau_{\max} = 100$. If μ_{rec} is not larger than $\tau_{\min}\mu_{\text{old}}$, we compute $\mu_{\text{new}} = 2\mu_{\text{last}}$. The GAMESS [30] package was used with the options SHIFT, DIIS and SOSCF set to true and other algorithms to false, except when SOSCF needed to be set to false due convergence failtures caused by instabilities of matrix manipulation (observed for distorted Cr_2 and CrC and for RhF_4^- in both geometries). Therefore, the dynamical level shifting of GAMESS is used in combination with DIIS and the SOSCF method is initiated when the orbital gradient falls bellow 0.25 a.u. (as default). This provides a combination of algorithms that should provide a strong convergence behavior and is useful to compare the present method with algorithms that are currently in use. A diagonalized Hamiltonian matrix was used as initial guess in all cases.

The results of the numerical tests are summarized in table 2. For the non-distorted geometries, both the DIIS algorithm and GAMESS behave very well, considering the number of iterations to achieve convergence. Convergence

Table 2
Comparison between DIIS, the unaccelerated TRRH, GAMESS and OD-DGTR.

Molecule	Method			
	DIIS	TRRH	GAMESS	OD-DGTR[a]
Non-distorted geometries				
Cr_2	10	30	18	13 (16)
CrC	30	200[b]	34	40 (62)
Rh_2	10	200[b]	16	17 (18)
RhF_4^-	38	200[b]	36	37 (44)
Li_9F_9	88	200[b]	200[b]	39 (227)
Rh complex	22	200[b]	44	38 (85)
Distorted geometries				
Cr_2	123	143	19	65 (83)
CrC	200[b]	150	104	45 (88)
Rh_2	200[b]	200[b]	200[b]	51 (97)
RhF_4^-	108	200[b]	100	76 (148)
Li_9F_9	200[b]	200[b]	200[b]	92 (182)
Rh complex	200[b]	200[b]	127	152 (249)

[a]The number in parenthesis is the number of Fock matrix evaluations. For the other methods, the of Fock matrix evaluations is equal to the number of iterations. [b]Maximum number of iterations (200) achieved.

was achieved in less than 40 iterations in all but the Li_9F_9 example by the DIIS method. The GAMESS calculation converged for all but Li_9F_9 example. The TRRH method is able to converge only for the Cr_2 example. Finally, as expected, convergence is achieved for all tests by the OD-DGTR algorithm. When comparing the number of iterations performed by DIIS versus OD-DGTR we see that, for these non-distorted geometries, the DIIS acceleration provides a faster convergence than OD-DGTR in most cases. The comparison is still more favorable to DIIS when the number of Fock matrix evaluations (the number between parentheses for OD-DGTR) is compared. For example, for the Li_9F_9 test, OD-DGTR performs 39 iterations, but required a total of 227 Fock matrix evaluations. The DIIS acceleration, on the other hand, requires only one Fock matrix evaluation per iteration. The fact that a lower number of iterations is performed by DIIS in these examples is not a surprise. For very well behaved tests, OD-DGTR will perform exactly as a classical fixed-point iteration. Therefore, its convergence will be clearly slower than the convergence of DIIS. For example, in the Rh_2 test, only one trust-region reduction was required and, therefore, 17 of the 18 iterations performed were simply fixed-point iterations. Of course, DIIS would have accelerated convergence in this case. The Li_9F_9 and Rh complex tests, however, must be considered apart. DIIS converged for these examples

in 88 and 22 iterations, respectively, while OD-DGTR converged in 39 and 38 iterations, but requiring several more Fock matrix evaluations. We note, however, that the convergence of DIIS is not guaranteed. Indeed, Thogersen et al. have published a result for the Rh complex in which they show that DIIS fails to converge. At the same time, the GAMESS calculation failed to converge in one of this examples even while the DIIS method was being applied. Obviously, the convergence of DIIS is highly dependent on the number of residual vectors used for extrapolation. The behavior of the combination of algorithms of the GAMESS package is somewhat better than of the OD-DGTR method given the number of iterations and Fock matrix evaluations for the cases for which it converges. We note, however, that the way that this combination of methods is used in GAMESS is optimized by experience, particularly the use of the SOSCF algorithm near the minimum and the heuristic for setting the level-shifting parameter at each iteration. On the other hand, the OD-DGTR method is applied as is, and is not dependent on user deffined parameters.

For the distorted geometries, the comparison of the convergence properties is quite different. Now, the DIIS acceleration is able to converge in only 2 of the 6 examples, and using more than 100 iterations in each case. The TRRH method converges for the Cr_2 and CrC examples, using 143 and 150 iterations, respectively. The GAMESS calculation converged for 4 of the 6 examples, being very fast for the distorted Cr_2 test, but using more than 100 iterations in the other three cases for which it converges. Finally, the OD-DGTR algorithm, although employing more iterations than in the non-distorted cases, converged again in all tests, as predicted by theory. The success of the optimally damped choice of the trust region must also be noted, since the convergence occurred in less than 100 iterations for all but the Rh complex example.

The convergence behavior of the DIIS, TRRH and OD-DGTR methods is represented in figure 3. An overall comparison of the convergence behavior shows that DIIS, as is known, has a quite unpredictable convergence, oscillating between several points even in the first iterations, even for problems for which it finally achieves convergence. The TRRH method, on the other side, starts with a very smooth convergence as seen in the first iterations, provided by the short step given by the small trust region defined by the choice of the level-shift parameter. However, when near the solution, convergence is not achieved and the algorithm oscillates between two or more points. This behavior was expected by the arguments given in the previous section (figure 1). Finally, the convergence of the OD-DGTR algorithm is smooth both far an near the solution of each problem. The energy decreases monotonically, as required by theory.

One should note that the convergence of the OD-DGTR method is faster than the convergence of the TRRH method, even in the first iterations. This is a consequence of the fact that each iteration of the OD-DGTR algorithm starts with an infinitely large trust region (a fixed-point iteration is performed) and the reduction of the trust region is given by the OD strategy. On the other hand,

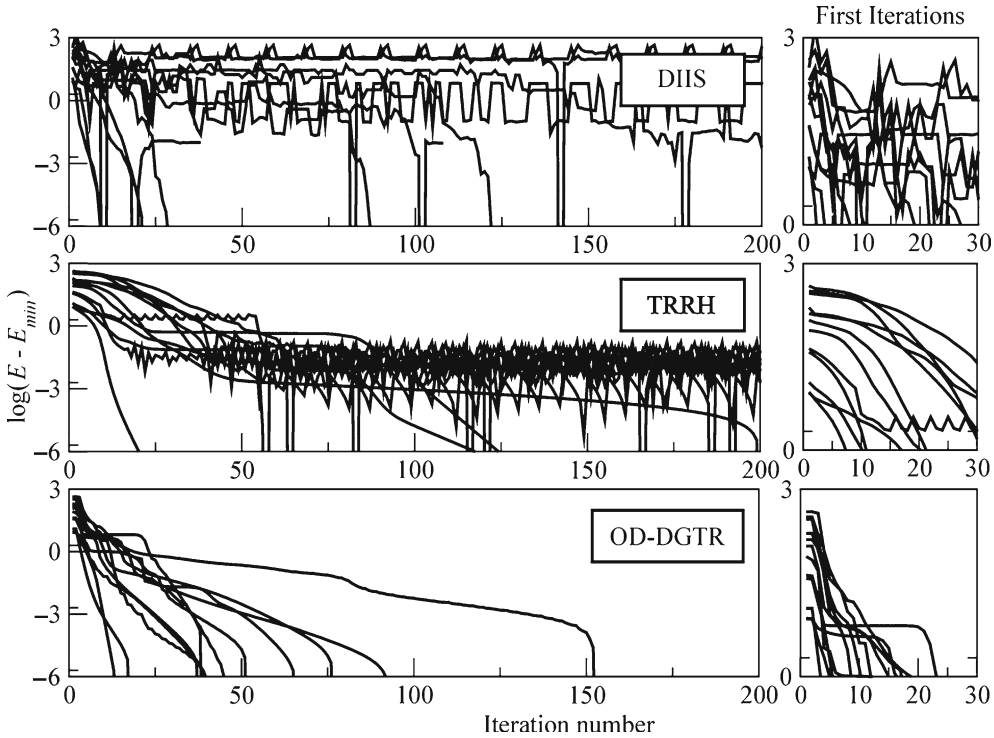

Figure 3. Comparison of the overall behavior of the three methods tested here. DIIS is very unpredictable and oscillates from the first iterations even if it finally converges. TRRH behaves smoothly when far from the solution, but oscillates when a minimum is approached. The OD-DGTR converges smoothly in all cases, and the optimally damped choice of μ provides a relatively fast energy decrease.

the current implementation of the TRRH method requires that the new density matrix is near to previous density matrix by at least 20%. This prevents the TRRH algorithm to have a faster convergence as these steps are too conservative, particularly when the algorithm is far from the solution.

A final comparison of the results is required for their full appreciation. table 3 shows the relative energy of the solution found by each method in each case. A zero indicate that the lowest energy solution was found with a precision of 10^{-9}. A positive value indicates how larger is the energy of the solution found by each method relatively to the best solution found. In well behaved cases, the DIIS method is the most successful method for finding lowest energy minima. It found the best solution for all but the CrC and RhF_4^- tests. The TRRH algorithm found the best solution in the case in which it converged. The GAMESS calculation found the best solution for 3 of the 5 tests for which it converged. Finally, the OD-DGTR algorithm found the best solution for 3 of the 6 examples and found energies larger than the best solution for other three

Table 3
Relative differences between the energies of the solutions and the lowest energy solution found by all algorithms.

Molecule	Method			
	DIIS	TRRH	GAMESS	OD-DGTR
Non-distorted geometries				
Cr_2	0	0	0	0
CrC	1.2×10^{-4}	–	0	1.2×10^{-4}
Rh_2	0	–	3.0×10^{-4}	0
RhF_4^-	3.6×10^{-6}	–	2.8×10^{-5}	0
Li_9F_9	0	–	–	5.6×10^{-6}
Rh complex	0	–	0	7.4×10^{-8}
Distorted geometries				
Cr_2	3.0×10^{-5}	1.2×10^{-4}	0	7.2×10^{-5}
CrC	–	0	1.1×10^{-4}	4.6×10^{-4}
Rh_2	–	–	–	0
RhF_4^-	5.8×10^{-6}	–	5.6×10^{-5}	0
Li_9F_9	–	–	–	0
Rh complex	–	–	0	3.3×10^{-5}

Zeroes indicate that the best point was found. Relative differences lower than 10^{-9} are considered zero.

cases, although in one of these cases the relative difference to the best solution found was less than 10^{-7}. The tendency of DIIS to converge to lowest energy solution comes from the fact that it can only converge to *aufbau* solutions and, therefore, it is very aggressive in providing convergence to the global solution.

For the examples with distorted geometries, the best solution was found by the GAMESS calculation for the Cr_2 and the Rh complex examples and by the TRRH algorithm for the CrC example. For the RhF_4^- test the best solution was found by the OD-DGTR method and, of course, the same happened for the Rh_2 and Li_9F_9 examples since this was the only method with which solutions were found. The comparison of these energies show that there is still a need for algorithms that strongly converge towards global minima, but that the OD-DGTR is competitive with current algorithms in this sense.

The numerical examples show the robustness of the globally convergent trust-region methods. Our previous GTR algorithm was also very robust, but it converged very slowly [5]. The use of density matrices as variables for the optimization in the context of trust-region methods allows one to define an algorithm with much higher efficiency, as was originally demonstrated by the TRSCF of Thogersen et al. [3]. This is a simple consequence of the the energy being quadratic as function of the density matrix. The use of the optimally damped strategy to compute the trust regions also provides the algorithm with higher efficiency.

8.1. Accelerations and non-monotone strategies

All the comparisons made for the behavior of the OD-DGTR algorithm are based on its pure implementation. The performance is good for problems with convergence instabilities. However, as was already mentioned, it is not better than the classical fixed-point algorithm if the fixed-point iteration provides a sufficient decrease of the energy at every iteration. Therefore, it is not recommendable for being used in routine calculations. How should accelerations or non-monotone strategies be incorporated to the DGTR framework without affecting the robustness of the algorithm? Any acceleration may be incorporated, provided that the accelerated point is feasible and the energy is at least as low as the energy of the current point. For example, after each iteration of the OD-DGTR algorithm, one could perform a DIIS extrapolation followed by a fixed-point iteration (providing feasibility to the accelerated point). If the energy at the accelerated point is lower than the energy at the current point, it can be accepted as a new iterate without affecting at all the convergence properties of the algorithm.

Non-monotone strategies can also be straightforwardly associated within the DGTR framework. The only requirement, again, is that after some iterations one obtains a feasible point with an energy at least as low as the energy of the lowest energy point found. For example, one could let DIIS to dictate the convergence with the sole requirement that, if after some number of DIIS iterations the energy has not decreased, one returns to the current point and computes a new trial point using the trust region strategy, until a lower energy density matrix is found.

This opens the possibility for the implementation within the DGTR framework of the novel and very effective accelerations schemes TRSCF and TRDSM introduced by Thogersen et al. [3,4] or the EDIIS scheme proposed by Kudin et al. [12], without affecting the robustness of the algorithm. Our claim is that this algorithm may be routinely used as a safeguard for convergence when oscillations or divergence are detected.

9. Conclusions and perspectives

In this paper we develop a theoretical framework of density based algorithms for SCF electronic structure calculations. The main features of the algorithm are a natural combination of previous developments on electronic structure calculations: The fixed-point iteration, the level-shifted equations and the optimal damped strategy. We prove that the solution of the level-shifted equations is the *global* minimization of a linear model of the energy in a smaller trust region defined by the level-shift parameter. The optimally damped choice of the trust region provides the algorithm with a nice practical behavior for solving HF

equations. Global convergence is obtained and, with the OD choice of the trust-region size, the number of iterations is reasonable, even for very difficult cases.

Why couldn't we call this a black-box algorithm yet? We believe the reasons are two: First, as was already mentioned, the convergence of the algorithm is not fast for well behaved cases, exactly resembling the fixed-point algorithm. Therefore, new studies for the definition of the best coupling of this method with accelerations and non-monotone strategies are required. Secondly, because we cannot guarantee convergence to global minimizers or to *aufbau* fixed points. No algorithm provides convergence to the global minimum and, therefore, unless some novel insight is obtained on the structure of this problem, convergence to global minimum will not be achieved by any means. On the other hand, some algorithms, when they converge, necessarily converge to *aufbau* fixed points (the fixed-point algorithm and DIIS, for example). This is somewhat desirable, since these solutions are interpreted as the electronic ground-state of the molecule under study. Again, for the development of a globally convergent algorithm that converges only to *aufbau* solutions, novel insights will be needed, particularly one would require a theoretical analysis of the type of functions for which global minima are *aufbau*. The definition of *aufbau* points given in section 3.1 may facilitate the access of the mathematical community to this problem.

Nevertheless, the OD-DGTR algorithm seems to be the first rigorously globally convergent method that has a nice practical behavior. We believe that the algorithm is ready to be implemented and tested in general purpose electronic structure packages for being used as an automatic safeguard for convergence.

Appendix A

A.1. Proof of theorem 1.

Define $A = D - \bar{D}$. By direct manipulation of matrices we see that

$$\mathrm{Tr}(ASAS) = \left\| S^{1/2} A S^{1/2} \right\|_F^2. \tag{A.1}$$

Therefore, the subproblem (10) is:

$$\text{Minimize } \mathrm{Tr}\left[2W(D - \bar{D}) \right] + \mu \left\| S^{1/2}(D - \bar{D})S^{1/2} \right\|_F^2 \tag{A.2}$$

subject to $D \in \mathcal{M}$.

For $\mu = 0$ the thesis follows as in theorem 3.1 of [5].

Let us change the variables in the following way:

$$Y = S^{1/2} D S^{1/2}, \quad Y^k = S^{1/2} \bar{D} S^{1/2}. \tag{A.3}$$

Then, the problem (A.2) can be reformulated as:

$$\text{Minimize } \mathrm{Tr}\left[2WS^{-1/2}(Y-Y^k)S^{-1/2}\right]+\mu\left\|Y-Y^k\right\|_F^2 \tag{A.4}$$

subject to $Y \in \mathcal{N}$, where

$$\mathcal{N}=\{D \in I\!R^{K\times K} \mid DD=D,\, D^T=D,\, \mathrm{Tr}(D)=N\}.$$

Define

$$G^k = S^{-1/2}WS^{-1/2}.$$

Since $\mathrm{Tr}(AB)=\mathrm{Tr}(BA)$ we get that (A.4) is equivalent to

$$\text{Minimize } \frac{2}{\mu}\mathrm{Tr}\left[G^k(Y-Y^k)\right]+\left\|Y-Y^k\right\|_F^2 \tag{A.5}$$

subject to $Y \in \mathcal{N}$.

Manipulating the objective function of (A.5) we obtain that this problem is equivalent to

$$\text{Minimize } \|Y-Z^k\|_F^2 \quad \text{subject to } Y \in \mathcal{N}, \tag{A.6}$$

where

$$Z^k = Y^k - \frac{1}{\mu}G^k. \tag{A.7}$$

Assume that $U\Sigma U^T$ is the spectral diagonalization of Z^k. Therefore, U is unitary and Σ is diagonal. Since $\|UA\|_F=\|AU\|_F=\|A\|_F$ for all A, we obtain that (A.6) is equivalent to

$$\text{Minimize } \|U^TYU-\Sigma\|_F^2 \quad \text{subject to } Y \in \mathcal{N}. \tag{A.8}$$

But $Y \in \mathcal{N}$ if, and only if, $U^TYU \in \mathcal{N}$. So, the solution of (A.8) can be expressed in the form

$$Y = UZU^T, \tag{A.9}$$

where Z solves

$$\text{Minimize } \|Z-\Sigma\|_F^2 \quad \text{subject to } Z \in \mathcal{N}. \tag{A.10}$$

Therefore (see justification later), assuming without loss of generality that the eigenvalues in the diagonal Σ are in increasing order:

$$Z=\begin{pmatrix}0 & 0\\ 0 & I_{N\times N}\end{pmatrix} \in I\!R^{K\times K}.$$

By (A.9), this implies that

$$Y = \bar{U}\bar{U}^T,$$

where the columns $\bar{U} \in \mathbb{R}^{K\times N}$ are eigenvectors associated with the N larger eigenvalues of Z^k. Therefore, by (A.3), the solution of (A.2) is given by

$$D = S^{-1/2} Y S^{-1/2} = V V^T,$$

where

$$V = S^{-1/2}\bar{U}. \tag{A.11}$$

The fact that the columns $\bar{U} \in \mathbb{R}^{K\times N}$ are eigenvectors associated with the N larger eigenvalues of Z^k implies that these columns are eigenvectors associated with the N smaller eigenvalues of $-\mu Z^k$. But

$$\begin{aligned} -\mu Z^k &= G^k - \mu Y^k \\ &= S^{-1/2} W S^{-1/2} - \mu S^{1/2}\bar{D}S^{1/2} \\ &= S^{-1/2}(W - \mu S\bar{D}S)S^{-1/2}. \end{aligned}$$

By (A.11) and the orthonormality of the columns of $\bar{U}$ we obtain (9). Moreover, by (A.11),

$$[S^{-1/2}(W - \mu S\bar{D}S)S^{-1/2}]\bar{U}_i = \lambda \bar{U}_i$$

is equivalent to $[W - \mu S\bar{D}S]V_i = \lambda_i S V_i$.

To finish the proof, let us compute Q_{opt}, the functional value of (10) at the solution D_{trial}:

$$\begin{aligned} Q_{\text{opt}} &= \text{Tr}[2W(D_{\text{trial}} - \bar{D}) + \mu \text{Tr}[(D_{\text{trial}} - \bar{D})S(D_{\text{trial}} - \bar{D})S] \\ &= 2\text{Tr}[WD_{\text{trial}}] - 2\text{Tr}[W\bar{D}] + \mu \text{Tr}(D_{\text{trial}}SD_{\text{trial}}S - D_{\text{trial}}S\bar{D}S - \bar{D}SD_{\text{trial}}S \\ &\quad + \bar{D}S\bar{D}S) \\ &= 2\text{Tr}[WD_{\text{trial}}] - 2\text{Tr}[W\bar{D}] + \mu \text{Tr}(D_{\text{trial}}SD_{\text{trial}}S) - 2\mu \text{Tr}(\bar{D}SD_{\text{trial}}S) \\ &\quad + \mu \text{Tr}(\bar{D}S\bar{D}S) \\ &= 2\text{Tr}[WD_{\text{trial}}] - 2\mu \text{Tr}(\bar{D}SD_{\text{trial}}S) - 2\text{Tr}[W\bar{D}] + \mu[\text{Tr}(D_{\text{trial}}SD_{\text{trial}}S) \\ &\quad + \text{Tr}(\bar{D}S\bar{D}S)] \\ &= 2\text{Tr}[WD_{\text{trial}}] - \mu\bar{D}SD_{\text{trial}}S] + c, \end{aligned}$$

where since $D_{\text{trial}}SD_{\text{trial}}S = D_{\text{trial}}S$ and $\text{Tr}(D_{\text{trial}}S) = N$,

$$\begin{aligned} c &= -2\text{Tr}[W\bar{D}] + \mu[\text{Tr}(D_{\text{trial}}SD_{\text{trial}}S) + \text{Tr}(\bar{D}S\bar{D}S)] \\ &= -2\text{Tr}[W\bar{D}] + \mu[\text{Tr}(D_{\text{trial}}S) + \text{Tr}(\bar{D}S\bar{D}S)] \\ &= -2\text{Tr}[W\bar{D}] + \mu[N + \text{Tr}(\bar{D}S\bar{D}S)]. \end{aligned}$$

Therefore,

$$\begin{aligned} Q_{\text{opt}} &= 2\text{Tr}[(W - \mu S\bar{D}S)D_{\text{trial}}] + c = 2\text{Tr}[(W - \mu S\bar{D}S)VV^T] + c \\ &= 2\text{Tr}[V\,Diag(\lambda_1, \cdots, \lambda_N)V^T] + c \\ &= 2(\lambda_1 + \ldots + \lambda_N) + c, \end{aligned}$$

as we wanted to prove. □

To complete the arguments that prove theorem 1, let us justify the given solution of (A.10).

Write $\Sigma = Diag(\lambda_1, \ldots, \lambda_K)$ and $Z = (z_{ij})$. Assume, without loss of generality, that $\lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_K$. Since $Z = Z^T$, the constraint $ZZ = Z$ is

$$z_{ij} = \sum_{k=1}^{K} z_{ik} z_{jk} \;\forall\, i, \quad j = 1, \ldots, K.$$

We define the relaxed problem

$$\text{Minimize } \|Z - \Sigma\|_F^2 \quad \text{subject to } Z = Z^T, \text{Tr}(Z) = N, z_{ii} = \sum_{j=1}^{K} z_{ij}^2, i = 1, \ldots, K. \tag{A.12}$$

The constraints of (A.12) are a subset of the constraints of (A.10). Therefore, if a global solution of (A.12) is feasible for (A.10), it will also be a global solution of (A.10).

Now, (A.12) may be written as:

$$\text{Minimize } \sum_{i=1}^{K} (z_{ii} - \lambda_i)^2 + \sum_{i=1}^{K} \sum_{j \neq i} z_{ij}^2$$

subject to $Z = Z^T$, $\text{Tr}(Z) = N$ and

$$z_{ii}^2 - z_{ii} + \sum_{j \neq i} z_{ij}^2 = 0, \quad i = 1, \ldots, N.$$

This is equivalent to

$$\text{Minimize } \sum_{i=1}^{K} z_{ii}^2 - 2 z_{ii} \lambda_i + \sum_{j \neq i} z_{ij}^2 \tag{A.13}$$

subject to $Z = Z^T$, $\sum_{i=1}^{K} z_{ii} = N$ and

$$z_{ii}^2 = z_{ii} - \sum_{j \neq i} z_{ij}^2, i = 1, \ldots, N. \tag{A.14}$$

Replacing (A.14) in (A.13) and simplifying, we obtain that the relaxed problem is equivalent to

$$\text{Minimize} \sum_{i=1}^{K}(z_{ii} - 2z_{ii}\lambda_i) \tag{A.15}$$

subject to

$$Z = Z^T, \sum_{i=1}^{K} z_{ii} = N, \tag{A.16}$$

and

$$z_{ii}^2 - z_{ii} = -\sum_{j \neq i} z_{ij}^2, \quad i = 1, \ldots, N. \tag{A.17}$$

Now, consider the linear programming problem

$$\text{Minimize} \sum_{i=1}^{K}(1 - 2\lambda_i)z_{ii} \tag{A.18}$$

subject to

$$Z = Z^T, \sum_{i=1}^{K} z_{ii} = N, \tag{A.19}$$

and

$$0 \leqslant z_{ii} \leqslant 1, \quad i = 1, \ldots, N. \tag{A.20}$$

A global solution Z_{glob} of (A.18)–(A.20) is such that $z_{ii} = 1$ for all $i = K, K - 1, \ldots, K - N + 1$ and $z_{ij} = 0$ for $i \neq j$. Since the constraint (A.17) implies the constraint (A.20) and Z_{glob} satisfies (A.17) it turns out that Z_{glob} is also a global solution of (A.15)–(A.17). So, Z_{glob} is a solution of (A.12). Since it is obviously feasible for (A.10), it is a global solution of (A.10).

A.2. Structure of the Li_9F_9 example

The structure of the Li_9F_9 example is provided in table A.1. The structure of the distorted Li_9F_9 corresponds to multiplying all coordinates in table A.1 by 2.

Table A.1
Structure of the Li_9F_9 example.

Atom	x	y	z
Li	0.00	3.00	3.00
Li	0.00	−3.00	3.00
F	0.00	0.00	3.00
Li	0.00	0.00	6.00
F	0.00	0.00	9.00
Li	0.00	0.00	12.00
F	0.00	0.00	15.00
Li	0.00	0.00	18.00
F	0.00	0.00	21.00
Li	0.00	0.00	24.00
F	0.00	0.00	27.00
Li	0.00	0.00	30.00
F	0.00	0.00	33.00
Li	0.00	0.00	36.00
F	0.00	0.00	39.00
Li	0.00	0.00	42.00
F	0.00	3.00	42.00
F	0.00	−3.00	42.00

Acknowledgments

We are indebted to Lea Thogersen, who provided us with the schematic code that computes μ in [4] We also thank Prof. Konstantin N. Kudin for providing the test example Li_9F_9. JBF was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Proc. 2004-13493-6 JMM supported by FAPESP and Conselho Nacional de Desenvolvimento Científico e Teconológico (CNPq), Grant 300017/1987-1. LM was supported by FAPESP, CNPq and by the Programa de Bolsas para Instrutores Graduados from UNICAMP.

References

[1] E. Cancès and C. Le Bris, Int. J. Quantum Chem. 79 (2000) 82.
[2] E. Cancès and C. Le Bris, ESAIM-Math. Model. Num. 34 (2000) 749.
[3] L. Thogersen, J. Olsen, D. Yeager, P. Jorgensen, P. Salek and T. Helgaker, J. Chem. Phys. 121 (2004) 16.
[4] L. Thogersen, J. Olsen, A. Köhn, P. Jorgensen, P. Salek and T. Helgaker, J. Chem. Phys. 123 (2005) 074103.
[5] J.B. Francisco, J.M. Martínez and L. Martínez, J. Chem. Phys. 121 (2004) 10863.
[6] G. Auchmuty and W. Jia, ESAIM-Math. Model. Num. 28 (1994) 575.
[7] T. Helgaker, P. Jorgensen and J. Olsen, *Molecular Electronic-Structure Theory* (John Wiley & Sons Inc., New York, NY 2000).
[8] G.B. Bacskay, Chem. Phys. 61 (1981) 385.

[9] A.D. Daniels and G.E. Scuseria, Phys. Chem. Chem. Phys. 2 (2000) 2173.
[10] P. Pulay, Chem. Phys. Lett. 180 (1991) 461.
[11] P. Pulay, J. Comput. Chem. 3 (1982) 556.
[12] K.N. Kudin, G.E. Scuseria and E. Cancès, J. Chem. Phys. 116 (2002) 8255.
[13] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, V.G. Zakrzewski, J.A. Montgomery, Jr., R.E. Stratmann, J.C. Burant, S. Dapprich, J.M. Millam, A.D. Daniels, K.N. Kudin, M.C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G.A. Petersson, P.Y. Ayala, Q. Cui, K. Morokuma, N. Rega, P. Salvador, J.J. Dannenberg, D.K. Malick, A.D. Rabuck, K. Raghavachari, J.B. Foresman, J. Cioslowski, J.V. Ortiz, A.G. Baboul, B.B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R.L. Martin, D.J. Fox, T. Keith, M.A. Al-Laham, C.Y. Peng, A. Nanayakkara, M. Challacombe, P.M.W. Gill, B. Johnson, W. Chen, M.W. Wong, J.L. Andres, C. Gonzalez, M. Head-Gordon, E.S. Replogle and J.A. Pople, Gaussian 03 (Gaussian, Inc., Pittsburgh PA, 2003).
[14] J.B. Francisco, Ph. D. Thesis, Department of Applied Mathematics, State University of Campinas (2005).
[15] A.R. Conn, N.I.M. Gould and Ph.L. Toint, *Trust-region Methods*, *MPS-SIAM Series on Optimization* (SIAM, Philadelphia, 2000).
[16] M.J.D. Powell, in: *Nonlinear Programming*, eds. J.B. Rosen, O.L. Mangasarian and K. Ritter (Academic Press, London, 1970).
[17] D.C. Sorensen, SIAM J. Numer. Anal. 19 (1982) 409.
[18] J.M. Martínez and S.A. Santos, Math. Program. 68 (1995) 267.
[19] R. Fletcher, *Practical Methods of Optimization*, 2nd ed. (Wiley, New York, 1987).
[20] P.L. Lions, Comm. Math. Phys. 109 (1987) 33.
[21] V.R. Saunders and I.H. Hillier, Int. J. Quantum Chem. 7 (1973) 699.
[22] L. Thogersen, Private Communication (2005).
[23] J. Barzilai and J.M. Borwein, IMA J. Numer. Anal. 8 (1988) 141.
[24] M. Raydan, IMA J. Numer. Anal. 13 (1993) 321.
[25] M. Raydan, SIAM J. Optimiz. 7 (1997) 26.
[26] F. Luengo, M. Raydan, W. Glunt and T.L. Hayden, Numer. Algorithms 30 (2002) 241.
[27] E.G. Birgin, J.M. Martínez and M. Raydan, SIAM J. Optimiz. 10 (2000) 1196.
[28] E.G. Birgin, J.M. Martínez and M. Raydan, IMA J. Numer. Anal. 23 (2003) 539.
[29] R. Fletcher, "On the Barzilai-Borwein method", Dundee, Scotland, 2001 (http://www.maths.dundee.ac.uk/~ftp/na-reports/NA207_RF.ps.Z).
[30] M.W. Schmidt, K.K. Baldridge, J.A. Boatz, S.T. Elbert, M.S. Gordon, J.H. Jensen, S. Koseki, N. Matsunaga, K.A. Nguyen, S.J. Su, T.L. Windus, M. Dupuis and J.A. Montgomery, J. Comput. Chem. 14 (1993) 1347.